Acesso Aberto a Dados de Pesquisa: levantamento de informações para a implantação de repositórios de dados

Rafael Port da Rocha[a], Leandro N. Ciuffo[b], Carolina H. Felicíssimo[b] ,

[a] Centro de Documentação e Acervo Digital (CEDAP), Universidade Federal do Rio Grande do Sul (UFRGS). Porto Alegre, Brasil
rafael.rocha@ufrgs.br

[b] Rede Nacional de Ensino e Pesquisa (RNP).
Rio de Janeiro, Brasil
{leandro.ciuffo,carolina.felicissimo}@rnp.br

Resumo. O valor estratégico do compartilhamento dos dados da pesquisa tem provocado grandes transformações nas instituições científicas no que diz respeito a prover meios para gerir dados digitais de pesquisa. Este artigo discute aspectos a serem considerados para a escolha de soluções tecnológicas para a implantação de repositórios de dados da pesquisa. Considera que critérios para avaliar softwares para repositório de dados devem ser elaborados com base em aspectos que envolvem princípios e modelo de referência para repositórios que são amplamente aceitos pela comunidade, como FAIR, Princípio de Citação e OAIS; e que esses critérios devem ser aplicados a partir da caracterização do ambiente do repositório, com a identificação dos produtores, dos consumidores e do ciclo de vida do dado. Apresenta aspectos a serem considerados na implantação de um repositório de dados de pesquisa, considerando FAIR, Princípio de Citação, OAIS e a caracterização do ambiente do repositório apoiada no ciclo de vida do dado. Apresenta três softwares mais usados para repositórios de dados, enfatizando características relacionadas aos aspectos discutidos. Propõe o desenvolvimento de critérios para avaliar softwares para repositórios de dados tendo como base os aspectos investigados no artigo. É parte da experiência do grupo GT-RDP Brasil na busca de soluções tecnológicas para o compartilhamento de dados.

Palavras Chave: Repositório de dados de pesquisa, Princípios FAIR, OAIS.
.
Eixo temático: Infraestrutura Tecnológica e Segurança

1 Introdução

O compartilhamento de dados da pesquisa é o ato de disponibilizar dados de pesquisa para outras pessoas para reutilização. Traz como benefícios permitir a reprodução ou a verificação da pesquisa; possibilitar a disponibilização ao público dos resultados de pesquisas financiadas com fundos públicos; permitir que outros façam novos questionamentos sobre os dados existentes e permitir avanços no estado da pesquisa e a inovação. [1]

Governos, comunidades e instituições nacionais e regionais passaram a identificar o valor estratégico desse compartilhamento, fomentando o desenvolvimento de infraestruturas e tecnologias que estimulassem a disponibilização dos dados. Atualmente, as nações mais desenvolvidas reconhecem essa necessidade, dando apoio a um conjunto de princípios para prover a abertura dos dados de pesquisa [2]. Agências e programas de fomento à pesquisa, como NSF e NIH nos EUA, Horizon2020 na Europa e mais recentemente a FAPESP no Brasil, estabelecem que projetos de pesquisa sejam acompanhados de planos de gestão de dados, que descrevem como os dados da pesquisa serão tratados durante o projeto e após sua conclusão. Instituições e organizações nacionais e internacionais movimentam-se no sentido de prover serviços de apoio ao compartilhamento e ao acesso aberto a dados de pesquisa, como serviços repositórios de dados de pesquisa.

Nesse cenário, a consolidação de estratégias, práticas, infraestruturas e tecnologias para gestão de dados ocorre por meio de uma comunidade internacional multidisciplinar, composta por pesquisadores, instituições, associações, coalizões e iniciativas, em sintonia com órgãos governamentais e instituições que promovem a pesquisa. A Research Data Alliance (RDA) e o World Data System (ICSU-WDS) são associações internacionais de destaque no apoio e no desenvolvimento de estratégias e soluções para o compartilhamento de dados.

O Grupo de Trabalho RDP-Brasil foi selecionado através de uma chamada de projetos sobre Acesso Aberto a Dados de Pesquisa (AADP) da Rede Nacional de Ensino e Pesquisa (RNP), em parceria com o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), para identificar práticas, mapear requisitos e prototipar um repositório que facilite o compartilhamento de dados científicos. O grupo é coordenado pela Universidade Federal do Rio Grande do Sul (UFRGS), com a participação da Universidade Federal do Rio Grande (FURG).

Esse artigo apresenta a investigação que está sendo realizada para atender a uma meta do GT-RDP Brasil, que compreende no levantamento comparativo dos serviços e soluções tecnológicas para compartilhamento de dados. O artigo discute e identifica aspectos a serem considerados para a realização deste levantamento, cujos resultados podem ser contribuições relevantes a instituições ou redes acadêmicas (NRENs) da América Latina que pretendem desenvolver projetos sobre Acesso Aberto a Dados de Pesquisa (AADP).

Os aspectos visam o desenvolvimento de um repositório para dados de pesquisa. A identificação desses aspectos parte das seguintes questões norteadoras:

- **Ambiente de Produção e Consumo dos Dados:** O repositório está inserido em um ambiente de produção e consumo dos dados que está focado em um ciclo de vida do dado da pesquisa.
- **Princípios e Modelos Norteadores:** Soluções para AADP estão se consolidando por meio de uma comunidade internacional multidisciplinar que atua na elaboração de princípios e modelos de referência que direcionam e orientam a tomada de decisões.

O artigo está estruturado da seguinte forma: a seção 2 apresenta aspectos a serem considerados na implantação de um repositório de dados, considerando FAIR, Princípio de Citação, OAIS, e a caracterização do ambiente do repositório sendo apoiada no ciclo de vida do dado. A seção 3 destaca três softwares mais usados para repositórios de dados, enfatizando características relacionadas aos aspectos discutidos. Nas considerações finais, são apresentados os próximos passos a serem desenvolvidos, a partir dos resultados obtidos.

2 Implantação de um Repositórios de Dados da Pesquisa

O modelo *Open Archival Information System* (OAIS) [3] serve como modelo de referência para o desenvolvimento de repositórios que estão comprometidos em preservar e manter acessível seus conteúdos digitais a longo prazo. É formado pelas especificações: **ambiente do repositório**, modelo de **representação das informações** a serem preservadas e **modelo funcional** do repositório. O ambiente de um repositório (figura 1) é composto pelo **produtor**, que fornece a informação a ser preservada, pelo **consumidor**, que interage com o ambiente para localizar e obter a informação preservada de seu interesse, e pela **administração**, que determina as políticas, controlando as responsabilidades de gestão. O modelo funcional (figura 1) compreende em entidades funcionais que compõem um repositório OAIS: ingestão, acesso, planejamento da preservação, armazenamento, gestão de dados e administração.

Para o desenvolvimento de um repositório, primeiramente é necessário compreender adequadamente o seu **ambiente**, isto é, identificar e caracterizar os **produtores** e os **consumidores** das informações. Isso irá determinar como serão os acordos de submissão, as estruturas para representação das informações, os recursos para acesso e para entrega da informação ao consumidor, incluindo informações de direitos de uso.

Em repositórios de dados da pesquisa, o **ambiente do repositório** está ligado ao ciclo de vida do dado, em que **produtores** são aqueles que desenvolvem ações de planejamento, coleta, processamento e análise de dados, atuando na fase ativa do desenvolvimento de uma pesquisa. Já os **consumidores** são aqueles que localizam e obtém os dados de pesquisa que foram disponibilizados para o compartilhamento, para produzir novos experimentos ou validar pesquisas. O repositório atua na recepção dos dados a serem preservados (ingestão), incluindo validação e verificação

de conformidades com acordos de submissão; no armazenamento a longo prazo dos dados, planejando e executando ações de preservação; e na disponibilização dos dados aos consumidores. Nesse contexto, os acordos de submissão versam em torno de aspectos que envolvem a preparação dos dados para fins de compartilhamento.

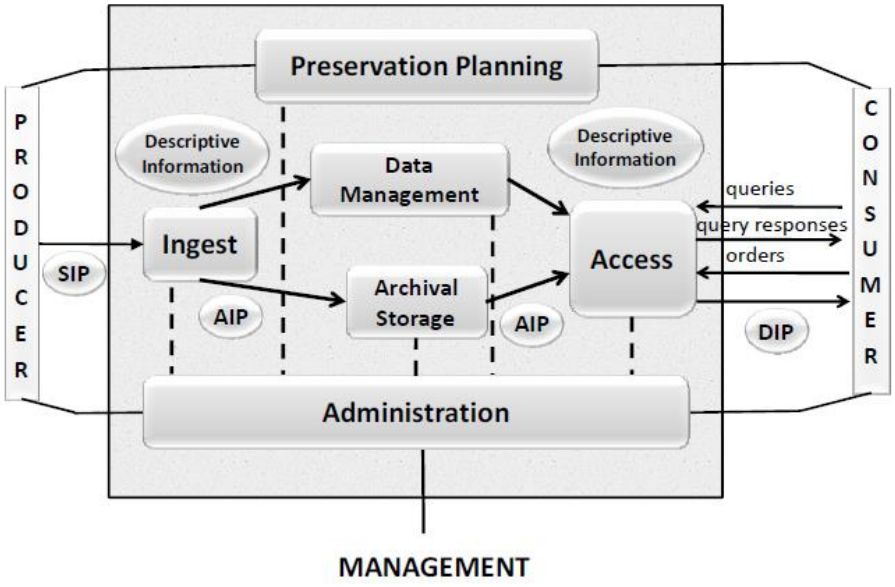

Fig. 1. Entidades Funcionais do Modelo OAIS. Fonte [3]

Para o levantamento comparativo de soluções para repositórios de dados da pesquisa, aspectos ligados ao **Ambiente de Produção e Consumo dos Dados** são obtidos a partir da análise do **ambiente** OAIS no contexto do ciclo de vida do dado da pesquisa. O texto a seguir apresenta as etapas do ciclo de vida do dado da pesquisa e aspectos dessas etapas que são relevantes na produção de dados da pesquisa para fins de compartilhamento em repositório OAIS. Para essa análise, foi adotado o ciclo de vida de dado da pesquisa do repositório UK Data Archive [4], composto pelas seguintes etapas:

- **Planejamento de Dados**: Desenhar pesquisa; planejar gestão de dados, consentir o compartilhamento, processamento, os protocolos de coleta e templates; explorar fontes de dados existentes.
- **Coleta de Dados**: Coletar dados; capturar dados com metadados; obter dados existentes de terceiros.
- **Processamento e Análise de Dados**: Entrar, transcrever e traduzir dado; checar, validar, limpar, anonimizar; derivar dados; descrever e documentar dado;

gerenciar e armazenar dado; analisar e interpretar dado; produzir saídas da pesquisa; citar fontes de dados.

- **Publicação e Compartilhamento de Dados**: Estabelecer *copyright*; criar documentação de usuário; criar metadados de descoberta; selecionar acesso apropriado aos dados; publicar/compartilhar dados; promover dados.
- **Preservação de Dados**: Migrar dados para o melhor formato/meio; armazenar e criar cópia de segurança dos dados; criar a documentação de preservação; preservar e curar dados.
- **Reuso de Dados**: Conduzir análises secundárias; realizar pesquisa de acompanhamento; conduzir revisões da pesquisa; examinar os achados; usar dados para ensino e aprendizado.

O estudo apresentado neste artigo inicia-se pela fase de **Publicação e Compartilhamento de Dados**, pois é nessa etapa do ciclo que os dados são preparados para o compartilhamento, seguindo acordos de submissão que obedecem a recomendações de compartilhamento de dados que estão sendo discutidas e propostas pela comunidade científica.

A fase de **Publicação e Compartilhamento de Dados**, no contexto de um repositório OAIS, implica em prover meios para permitir que os **consumidores** descubram os dados, tenham acesso aos dados e disponham de informações suficientes que permitam que esses dados sejam adequadamente usados, seguindo os princípios legais e em conformidade com aquilo que foi acordado com o **produtor** dos dados. Além disso, o repositório deve prover meios para que o usuário dos dados possam citar adequadamente esses dados, de forma que não fiquem dúvidas quanto a sua proveniência, incluindo referência ao uso de versões ou subconjuntos de conjuntos de dados. A **Publicação e Compartilhamento de Dados** requer que os repositórios observem aspectos como: metadados, proveniência, contratos e licenças, restrições de acesso, e recursos para a citação dos dados.

A citação de dados é mais ampla que a citação de publicações científicas. A citação dos dados "abre a questão da terminologia usada para descrever referências mais granulares aos dados, incluindo subconjuntos de observações, variáveis ou outros componentes, assim como subconjuntos de um conjunto maior de dados. Essas referências granulares são frequentemente necessárias no texto para descrever o suporte probatório preciso para uma tabela de dados, figura ou análise, e são análogos à "citação de citação" usada na profissão jurídica ou na "referência de página" usada na citação de um periódico [5]".

Pesquisas podem envolver versões, partes temporais ou subconjuntos de dados e a citação destes dados deve identificar essas questões. Force 11 [6] apresentou um conjunto de princípios que cobrem o propósito, a função e os atributos da citação, que são: **Importância**, pois dados são produtos citáveis, ou seja, legítimos da pesquisa; **Crédito e Atribuição,** aos que contribuíram na formação dos dados; **Evidência**, pois dados devem ser citados na literatura quando uma afirmação baseia-se neles;

Identificação Única, global, compreensível por máquina e amplamente usada; **Acesso,** aos dados e aos seus metadados, sua documentação e materiais que permitam seu uso por máquinas e por humanos; **Persistência**, com identificadores e metadados que persistem mesmo quando os dados tornam-se indisponíveis; **Especificidade e Verificabilidade,** pois citações ou metadados de citação devem incluir informações de proveniência e fixidez suficientes para facilitar a verificação de que fatia de tempo, versão ou parte granular dos dados obtidos subsequentemente é a mesma que a que foi originalmente citada; e **Interoperabilidade e Flexibilidade**, combinando flexibilidade para acomodar práticas variantes entre as comunidades, mas não sendo diferentes o suficiente para comprometer a interoperabilidade das práticas de citação entre as comunidades.

A questão da **proveniência** dos dados a serem armazenados liga os repositórios dados da pesquisa às fases de **Coleta de Dados** e de **Processamento e Análise de Dados**, do ciclo de vida do dado, pois são nelas que são geradas e estruturadas as informações de proveniência que deverão ser armazenadas junto aos dados no repositório. Embora a obtenção dessas informações seja de responsabilidades dos **produtores**, repositórios precisam compreender a comunidade, a fim de estabelecer políticas e metadados para proveniência, assim como estruturas para representar os dados.

Proveniência dos dados compreende em prover "informações sobre entidades, atividades e pessoas envolvidas na produção de um dado ou coisa, que podem ser usadas para formar avaliações sobre sua qualidade, confiabilidade ou fidedignidade". [7] . Isso implica em prover instrumentos para documentar como os dados foram produzidos, que envolvem gerenciar as versões dos dados, identificar unicamente os conjuntos de dados (incluindo suas versões ou subconjuntos), e até mesmo registrar automaticamente o fluxo da produção dos dados, quando os mesmos são produzidos a partir de sistemas de workflow [8].

Para auxiliar a **Publicação e o Compartilhamento de Dados** da pesquisa, os princípios FAIR[1] foram propostos e estão sendo amplamente discutidos pela comunidade que pesquisa e promove dados de pesquisa. Prover meios para o compartilhamento de dados que seguem os princípios FAIR é um aspecto importante a ser considerado no desenvolvimento de um repositório de dados no cenário atual, pois os mesmos visam incrementar a encontrabilidade, a acessibilidade, a interoperabilidade e o reuso de ativos digitais, com enfoque na acionabilidade por máquinas, buscando a menor intervenção humana possível.

Os princípios FAIR [9] indicam que dados devem ser localizáveis (*Findable*), acessíveis (*Accessible*), interoperáveis (*Interoperable*) e Reusáveis (*Reusable*). Para dados serem FAIR, estes devem ser atribuídos a identificadores únicos, persistentes e

[1] Findable, Accessible, Interoperable, Reusable. https://www.go-fair.org/fair-principles/

globais (F). Devem ser descritos por metadados indexáveis e ricos (F), representados em linguagens formais (I), aceitos pela comunidade (R), com atributos relevantes, precisos e úteis ao contexto (R), incluindo metadados de proveniência (R) e usando vocabulários controlados que seguem princípios FAIR (I). Estes dados devem ser recuperáveis pelo seu identificador através de um protocolo de comunicação padronizado, aberto, gratuito (A). Também devem ser acompanhados de licenças claras e acessíveis (R), e referências qualificadas devem ligar (meta)dados para enriquecer o conhecimento sobre os mesmos (I).

A fase de **Preservação dos Dados** envolve o armazenamento dos dados em um repositório digital confiável. Repositório confiável é aquele "cuja missão é fornecer acesso de longo prazo a recursos digitais gerenciados; que aceita a responsabilidade pela manutenção a longo prazo dos recursos digitais em nome de seus depositantes e em benefício dos usuários atuais e futuros; que projeta seu(s) sistema(s) de acordo com as convenções e os padrões comumente aceitos para garantir o gerenciamento, acesso e segurança contínuos dos materiais depositados nele; que estabelece metodologias para avaliação de sistemas que atendem às expectativas de confiabilidade da comunidade; que cumpre suas responsabilidades com depositantes e usuários de forma aberta e explícita; e cujas políticas, práticas e desempenho podem ser auditados e medidos." [10]

Para serem confiáveis, repositórios devem atender a critérios técnicos de gerenciamento do objeto digital, assim como relacionados à infraestrutura organizacional necessária para gerir o repositório a longo prazo. Esses critérios surgem do modelo OAIS e o atendimento a eles é verificado através do uso de instrumentos de certificação de repositórios confiáveis, como ISO 16363, Data Seal of Approval[2] e CoreTrustSeal[3].

Os critérios de infraestrutura administrativa servem para certificar que o repositório tem condições administrativas para funcionamento a longo prazo, isto é, que repositório possui governança, transparência, estrutura organizacional e de pessoal, e sustentabilidade financeira, além de ser capaz de gerenciar licenças e contratos.

Critérios para gestão do objeto digital envolvem aspectos técnicos ligados aos componentes funcionais (figura 1) e de representação da informação do modelo OAIS, com grande parte de suas funções sendo absorvidas pelo software do repositório. Envolve a definição das estruturas dos pacotes para representar a informação a ser preservada, incluindo o objeto conceitual, suas representações digitais, informações de representações e metadados descritivos, técnicos, estruturais e administrativos. Envolve também funções de ingestão (verificação e validação das

[2]Data Seal Approval. https://www.datasealofapproval.org/en/
[3]Core Trustworthy Data Repositories. https://www.coretrustseal.org

informações submetidas, considerando especificações e procedimentos planejados), de armazenamento (verificando e mantendo integridade dos materiais, mídias), de acesso (busca, restrições de acesso e entrega ao consumidor), de planejando da preservação e de gestão de metadados.

O uso de microserviços de curadoria digital é uma estratégia para a automatização, em repositórios, de procedimentos ligados à gestão do objeto digital [11]. Microserviços operam em objetos submetidos e armazenados no repositório em uma configuração similar a uma linha de produção, executando tarefas como definir de identificador persistente ao objeto submetido, criar *checksum* (fixidez), verificar a integridade, verificar a conformidade dos pacotes com as especificações planejadas, caracterizar os formatos dos arquivos submetidos, validar dos formatos, extrair metadados técnicos dos objetos submetidos, verificar a ocorrência de vírus, entre outros. Arquivematica é um exemplo de software que disponibiliza microserviços para curadoria digital [12].

Os aspectos discutidos nessa seção aplicam-se à implantação de um repositório, e não à avaliação de um software de forma isolada. Partem de um pressuposto que o ambiente do repositório é conhecido, isto é, que são caracterizados os produtores e os consumidores dos dados, os tipos de dados produzidos, assim como o ciclo de vida do dado de pesquisa e as formas de coleta, processamento e análise dos dados. A seguir uma compilação dos aspectos discutidos nessa seção é apresentada. Essa compilação está focada nos princípios e modelo de referência FAIR[9], Citação[6] e OAIS [3]:

- **Proveniência**: Prover meios para documentar como os dados foram produzidos, via metadados e documentos; a partir de requisitos que consideram as etapas de planejamento, coleta, processamento e análise, considerando que os dados em questão podem ser versões, séries, subconjuntos, com coleta automatizada, incluindo fluxos de coleta, etc.
- **Citação**: Prover meios para permitir a citação dos dados, via metadados de citação e identificadores persistentes que levam a esses metadados e a outros documentos. Atribuindo créditos aos produtores, disponibilizando informações de proveniência e que permitem a verificação de fixidez, considerando que os dados em questão podem ser versões, séries, subconjuntos, etc .
- **Metadados**: prover meios para produzir, representar e gerenciar metadados ricos, precisos, indexáveis, úteis ao contexto, aceitos pela comunidade e compreensíveis por máquinas. Que permitam a descoberta e uso dos dados (metadados descritivos), a obtenção de informações de proveniência, de direitos de uso, sobre características técnicas dos objetos digitais (metadados técnicos), e sobre as ações de curadoria digital que foram realizadas (metadados administrativos/preservação digital). Incluindo a descrição dos elementos que compõem o objeto digital e suas estruturas (metadados estruturais) .
- **Licenças**: prover meios para gerenciar licenças e embargos, a fim de facilitar a gestão do repositório.
- **Representação do Objeto Digital**: prover meios para representar os dados (pacotes), considerando suas versões, suas diversas representações (em diversos formatos), seus metadados e documentos associados; com identificação única,

persistente e global, considerando versões e subconjuntos; e cujas estruturas de representação possam ser verificáveis por máquinas a partir de especificações planejadas.

- **Submissão da Informação**: prover meios para submissão dos dados ao repositório, observando fluxo de submissão e funções (muitos podendo ser implementados como microserviços de curadoria) que verificam se o material submetido está íntegro e em conformidade com as especificações planejadas (estrutura do pacote, metadados, proveniência, planejamento da preservação).
- **Armazenamento**: prover meios para armazenamento seguro da informação, em pacotes de armazenamento em conformidade com as especificações planejadas; para checagem da integridade das informações; com o registro das ações de preservação digital e seus efeitos e o armazenamento de objetos digitais decorrentes dessas ações (como novas representações decorrentes de migrações).
- **Acesso**: prover meios para descoberta dos dados; restringir o acesso a dados a pessoas ou grupos autorizados; entregar dados ao consumidor em formatos usados por estes; prover acesso aos dados, seus metadados e outras informações, como proveniência e licenças, através de protocolos de comunicação padronizados, abertos e gratuitos.
- **Planejamento da Preservação**: prover meios para que o planejamento da preservação trabalhe de forma integrada com os serviços oferecidos pelo repositório, como por exemplo, gerenciar formatos, monitorar formatos, microserviços de curadoria digital.

Conhecer os produtores e os consumidores dos dados implica em identificar configurações que as pesquisas assumem para essa comunidade com relação aos dados. Observamos que muitas comunidades de pesquisa movem-se para uma configuração que é caracterizada por usar dados intensivamente (*data-intensive*), por envolver áreas multidisciplinares, atuar em rede, de forma colaborativa e distribuída [13]. Repositórios com essa configuração normalmente armazenam grandes conjuntos de dados; são mais propensos a adotar poucos formatos, unificados e padronizados, assim como fluxos e processos de coleta e análise, que muitas vezes são automatizados. Um exemplo é o caso da área das ciências da vida, que desenvolveu a Arquitetura ISA[4], para facilitar a coleta, a curadoria, o gerenciamento e a reutilização de conjuntos de dados em conformidade com os padrões.

Oposto aos repositórios que armazenam grandes conjuntos, com relativamente pouca diversidade e quantidade, temos os repositórios voltados à cauda longa dos dados. Esses repositórios armazenam uma grande quantidade de conjuntos de dados, com dimensões menores, que assumem uma grande diversidade de configurações. [14] apresentam características que diferenciam dados da cauda longa e dados do topo da curva:

- Dados da Cauda Longa: Heterogêneo, pequeno, padrões exclusivos, não regulados, curadoria individual, armazenados em repositório institucional, geral ou nenhum repositório.
- Dados do Topo da Curva: Homogêneos, grandes, padrões comuns, regulamentado, curadoria central, armazenados em repositórios disciplinares.

[4] ISA - http://www.isacommons.org/

3 Software para Repositórios de Dados da Pesquisa

Atualmente vários softwares livres são usados para repositórios de dados da pesquisa. São softwares que foram desenvolvidos especificamente para dados de pesquisa (como Dataverse) ou softwares que originalmente foram desenvolvidos para outros propósitos, como para repositório institucional (como DSpace) e para prover a abertura de dados governamentais (como CKAN). Tendo como referência o diretório de dados da pesquisa Res3Data.org, Dspace, Dataverse e CKAN são os softwares mais usados.

Estes softwares oferecem uma solução integrada, isto é, suas funcionalidades abrangem (em maior ou menor grau) as entidades funcionais do modelo OAIS (figura 1): submissão, armazenamento a longo prazo, acesso, gestão de metadados, administração e preservação digital. A vantagem dessa solução é que tudo está presente em um único software. Traz facilidades para instalação e operação (integrada), e proporciona robustez do ambiente. A desvantagem está na dificuldade em adaptar o software às características do repositório, isto é, para estender o software para que este passe a atender a aspectos do ambiente do repositório não contemplados pelo software original.

Outra solução que vem sendo adotada por alguns repositórios é implementar componentes funcionais de OAIS na forma de serviços independentes, desenvolvidos a partir do reuso de diversos tipos de software, que interagem entre si para atender às funções do repositório. Esse tipo de solução oferece maior flexibilidade de adaptação às necessidades do repositório, principalmente quando estes necessitam armazenar dados com características especiais ou que são coletados de forma automática.

EUDAT [15] é um exemplo solução com essas características, que apresenta uma infraestrutura com serviços para replicar dados (B2Safe), para computar dados (B2Stage), para localizar dados (B2Find), para armazenar, compartilhar e publicar dados (B2Share), para definir identificadores globais e persistentes (B2Handle) e para sincronizar e trocar dados (B2Drop). EUDAT usa vários softwares na implementação desses serviços, como por exemplo, iRods[5], que é usado para o armazenamento a longo prazo (B2Safe), Ivenio[6], usado para armazenamento (B2Safe), usado CKAN para busca (B2Find), Handle System, para identificação (BD2HJandle), entre outros [16]. Archivemarica [12] é um software que se enquadra nessa categoria, pois é focado em prover microserviços de ingestão, transferindo os pacotes ingeridos (verificados e validados) para ambientes de acesso e de armazenamento a longo prazo.

Os aspectos apresentados na seção anterior objetivam identificar as características do repositório de dados, considerando princípios e modelo de referência FAIR, Citação e OAIS. Conhecidas as características do repositório, soluções tecnológicas são buscadas para atender o funcionamento do repositório, no qual o software atua na automatização de funções do modelo OAIS.

---

[5] iRods - Software para gerenciar armazenamento seguro, replicado, baseado em regras e microserviços, https://irods.org

[6] Framework para construção de ambiente para biblioteca e rpositório digital, https://invenio-software.org/

Repositórios de cauda longa normalmente demandam por requisitos mais genéricos, a fim de atender a diversidade da cauda longa dos dados, sendo mais fácil o uso de softwares integrados, como DSpace, KCAN e Datavere. Já em casos de repositórios voltados a áreas específicas, que possuem particularidades para metadados, representações de dados, procedimentos de coleta, produção e processamento, pode ser mais conveniente a adoção a estratégia de implementar componentes OAIS como serviços.

Dataverse [17] é um software integrado para publicação, compartilhamento e armazenamento de dados. Traz facilidades para representar cenários que são compostos por diversas entidades hierárquicas (como universidades, unidades ou grupos), que são autônomas, isto é, que têm poder para definir quem pode criar, autorizar a publicação ou acessar conjuntos de dados, estabelecer licenças e definir que o uso dos dados somente pode ser feito mediante solicitação. Também permite a configuração e uso de diversos esquemas de metadados, gerencia versões de conjuntos de dados, identifica unicamente conjuntos de dados (considerando versões) de forma universal e persistente (sistemas DOI ou Handle System), disponibiliza metadados de citação e uma estrutura para citação [18] que envolve a verificação da fixidez do material citado. Permite o armazenamento de documentos complementares junto a conjunto de dados, a adição de ferramentas de análise de dados, a customização de interfaces, o uso de serviços de caracterização de formatos, cópias de segurança compartilhada com outros repositórios (Lockss[7]), submissão por máquinas (Sword) e colheita de metadados (OAI-PMH[8]). É usado por instituições (heiData[9]/Univ. Heidelberg), por grupos de instituições (DataverseNL[10]/Holanda, Abacus[11]/Canada, Texas Digital Library[12]/EUA), e para repositórios temáticos (ICRIAT[13]/Agricultura) e multidisciplinares (Australian Data Archive[14]). Alguns repositórios que usam Dataverse são repositórios digitais confiáveis certificados, como Australian Data Archive e TiU Dataverse[15]/DatavereNL, demonstrando que o software atende a necessidades desse tipo. Dataverse atende a requisitos FAIR , como demonstrado por [9], que relaciona funcionalidades acima citadas com princípios FAIR.

DSpace é um software desenvolvido para repositório institucional. Assim como Dataverse, permite a representação de unidades e subunidades autônomas, com a configuração de fluxos específicos de submissão, uso de diversos esquemas de metadados, identificação universal e persistente através de Handle System, cópias distribuídas de segurança (Lockss), e submissão por máquina (Sword) e colheita de

---

[7] Lots of Copies Keep Stuff Safe. https://www.lockss.org/
[8] OAI-Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/
[9] heiDATA. https://heidata.uni-heidelberg.de/
[10] DataverseNL https://dataverse.nl/
[11] Abacus Dataverse Network https://abacus.library.ubc.ca
[12] TDL/Texas Data Repository http://data.tdl.org/
[13] Int. Crops Research Institute for the Semi-Arid Tropics- http://dataverse.icrisat.org/
[14] Australian Data Archive. https://dataverse.ada.edu.au/
[15] TiU - CoreTrustSeal https://www.coretrustseal.org/wp-content/uploads/2018/04/Tilburg-University-Dataverse.pdf

metadados (OAI-PMH). É usado por reconhecidos repositórios de dados, como o repositório multidisciplinar Dryad[16] e o repositorio institucional DataShare[17]. DataShare[18] é um exemplo de repositório confiável, certificado, que usa DSpace, demonstrando que repositórios em DSpace podem ser certificados como confiáveis.

Como Dataverse foi desenvolvido para repositório de dados, a representação e a gestão automatizada dos conjuntos de dados é estruturada através do conceito Dataset, que inclui dados, metadados de citação, metadados específicos, documentação adicional, citação, gerenciamento de versões, etc. Já DSpace está estruturado no conceito de coleção de itens, com cada item sendo compostos por pastas (bundles) que contém arquivos (bitstreams). No caso do uso de DSpace para gerenciar dados, é necessário configurar metadados, fluxos e interfaces de usuário para conduzir a submissão de dados, como feito no repositório DataShare [19]. DSpace não gerencia versões, incluindo identificação e citação de conjuntos versionados, como Dataverse.

CKAN é um software desenvolvido para abertura de dados governamentais. Nele, conjuntos de dados pertencem a organizações. Membros de organizações são autorizados a acessar, tornar público, editar e remover conjuntos de dados. CKAN possui um pequeno conjunto próprio de metadados, e permite a inclusão de campos para descrever um conjunto de dados. Permite também a seleção de licenças, controla versões de conjuntos de dados, mantém histórico das modificações e gerencia formatos. Não possui identificador global persistente, nem permite o uso de esquemas de metadados. Mas oferece grande flexibilidade para a criação de extensões, já havendo uma extensão para identificador persistente (DOI) e uma para metadados estruturados. Em repositório de dados, o conceito organização de CKAN é usado para representar instituições, unidades ou grupos. Conjuntos de dados podem conter vários arquivos, representando dados e informações adicionais.

CKAN é mais limitado que Dataverse e DSpace no que diz respeito a recursos para organizar unidades e seus dados, para definir políticas de gestão e submissão de dados específicas para cada unidade, assim como no que diz respeito a metadados e identificador persistente. CKAN é uma boa solução quando usado como ambiente de publicação e acesso a dados, em que a submissão é feita por outro sistema. É o que ocorre no repositório de dados da Universidade de Bristol [19], no qual pesquisadores solicitam espaço de armazenamento e submetem os dados através do sistema de gerenciamento de pesquisa da universidade. Esse sistema conduz a avaliação dos dados e os publica via CKAN.

Considerando os três softwares, Dataverse possui recursos para configuração de vários tipos de ambientes de repositório, incluindo hierarquias organizacionais, políticas de gestão distintas para unidades ou grupos, incluindo esquemas de metadados e licenças. Isso é possível em DSpace, entretanto exige adaptações configurações, com algumas limitações no controle de versões. CKAN já é mais limitado, entretanto é uma boa alternativa quando usado como serviço de publicação e acesso, com a submissão e preservação digital sendo realizados por outros ambientes.

---

[16] Dryad - https://datadryad.org/
[17] Edinburgh DataShare. https://datashare.is.ed.ac.uk/
[18] Datashare - Data Seal Approval - https://assessment.datasealofapproval.org/assessment_175/seal/pdf/

Considerações Finais

Este artigo apresentou resultados preliminares do projeto RDP-Brasil no que diz respeito à meta que compreende no levantamento comparativo dos serviços e soluções tecnológicas para compartilhamento de dados. Enfatizou que o desenvolvimento de um repositório parte da compreensão do ambiente do repositório, com a caracterização dos produtores, dos consumidores, das estratégias de coleta, do ciclo de vida do dado, assim como de representação dos dados e dos metadados. Apresentou aspectos a serem considerados na implantação de um repositório de dados, considerando princípios estabelecidos e aceitos pela comunidade (FAIR e Citação) e o modelo de referência OAIS. Destacou que a implantação de um repositório deve considerar avaliar aspectos de proveniência, citação, metadados, licença, representação do objeto digital, submissão da informação, armazenamento, acesso e planejamento da preservação.

Identificou e apresentou características dos três softwares mais usados para repositório de dados, enfatizando aspectos como identificação de objeto, controle de versão, metadados, configuração de unidades com autonomia e estruturas próprias para gerenciar e representar conjuntos de dados e metadados, entre outras.

Vários estudos têm sido desenvolvidos no sentido de auxiliar a escolha de soluções tecnológicas para repositórios de dados e de dados de pesquisa (como [20] , [21], [22] e [23]). Esta investigação distingue-se desses estudos em dois aspectos: (i) parte da premissa que os critérios devem ser aplicados após a identificação das características do ambiente do repositório, incluindo o ciclo de vida dos dados; (ii) e da premissa que os critérios devem ser elaborados com base em aspectos que envolvem princípios e modelo de referência amplamente aceitos pela comunidade, como FAIR, Princípio de Citação e OAIS.

Como próximas ações, serão elaborados critérios para caracterizar o ambiente de um repositório, e para avaliar comparativamente soluções tecnológicas para esse ambiente, tendo como base os aspectos discutidos nesse artigo. Então serão investigadas soluções tecnológicas para um repositório de dados de pesquisa.

Referencias

1. Borgman, Christine L.: The conundrum of sharing research data. Journal of the Association for Information Science and Technology, vol. 63, n.6, 1059 a 1078 (2012). https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22634
2. Foreign & Commonwealth Office: G8 Science Ministers Statement. Foreign & Commonwealth Office (2013). https://www.gov.uk/government/news/g8-science-ministers-statement

3. Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System. CCSDS, Washington (2012). https://public.ccsds.org/pubs/650x0m2.pdf
4. The UK Data Service: Research data lifecycle, www.data-archive.ac.uk/create-manage/life-cycle
5. CODATA-ICSTI Task Group on Data Citation Standards and Practices:Out of Cite, Out of Mind - The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, vol. 12 (2013). https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf
6. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11 (2014) https://doi.org/10.25490/a97f-egyk
7. Groth, P., Moreau, L.: PROV-Overview - An Overview of the PROV Family of Documents. W3C Working Group, Note 30 (2013). https://www.w3.org/TR/prov-overview/
8. Bechhofer, S., et al.: Research Objects - Towards Exchange and Reuse of Digital Knowledge. Nature Precedings (2010). http://dx.doi.org/10.1038/npre.2010.4626.1
9. Wilkinson, M. D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, vol. 3 (2016). https://www.nature.com/articles/sdata201618
10. Jantz, R., Giarlo, M.: Digital Preservation - Architecture and Technology for Trusted Digital Repositories. Dlib Magazine, vol. 11, n. 6 (2006) http://www.dlib.org/dlib/june05/jantz/06jantz.html
11. Abrams, S. et al.: Curation Micro-Services: A Pipeline Metaphor for Repositories. Journal of Digital Information, vol 12, n.2 (2011) https://journals.tdl.org/jodi/index.php/jodi/article/view/1605/1766
12. van Garderen, P.: Archivematica - using micro-services and open-sou rce software to deliver a comprehensive digital curation solution. En 7th International Conference on Preservation of Digital Objects, iPRES (2010) www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf
13. Tenopir, C et al.: Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS ONE, vol. 10, n. 8 (2015) https://doi.org/10.1371/journal.pone.0134826
14. e-IRG: Long Tail of Data. e-IRG Task Force Report. (2016). http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf
15. Lecarpentier, D. et al.:EUDAT: A New Cross-Disciplinary Data Infrastructure for Science . International Journal of Digital Curation, vol. 8, n.1 (2013) http://www.ijdc.net/article/view/8.1.279
16. EUDAT. EUDAT Primer. https://eudat.eu/services/userdoc/eudat-primer
17. Crosas, M.: The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine, vol 17, n.1/2 (2011) . http://www.dlib.org/dlib/january11/crosas/01crosas.html
18. Altman, M.; King, G.: A Proposed Standard for the Scholarly Citation of Quantitative Data. D-Lib Magazine, vol 13, n.3/4 (2007). http://www.dlib.org/dlib/march07/altman/03altman.html
19. University of Edinburgh: Edinburgh DataShare: Depositor's User Guide (2018) . https://www.ed.ac.uk/files/atoms/files/datashare-january2018.pdf

20. Amorim, Ricardo C. et al. A comparison of research data management platforms: architecture, flexible metadata and interoperability. Univ Access Inf Soc (2017) 16: 851. https://doi.org/10.1007/s10209-016-0475-y
21. Lewis, John A.: Research Data Management Technical Infrastructure: A Review of Options for Development at the University of Sheffield. University of Sheffield (2014). https://figshare.com/articles/Research_Data_Management_Technical_Infrastructure_A_Review_of_Options_for_Development_at_the_University_of_Sheffield/1202230
22. Bankier,J., Gleason, K. Institutional repository software comparison. UNESCO (2014). http://www.unesco.org/ulis/cgi-bin/ulis.pl?catno=227115
23. Pyrounakis, G.; Nikolaidou, M.: Comparing Open Source Digital Library Software. Handbook of Research on Digital Libraries (2009). https://www.researchgate.net/publication/247933909_Comparing_Open_Source_Digital_Library_Software